Assignment 10: Data Scraping

Aileen Lavelle

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A10_DataScraping.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
library(tidyverse)
library(lubridate)
library(rvest)
library(cowplot)
library(patchwork)
library(here); here()
```

```
## [1] "/Users/aileen/Desktop/Duke/Environmental_Data_Analytics/EDA_Spring_2023_corrected"
```

```
Aileentheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        axis.ticks = element_line(color = "black"),
        plot.background = element_rect(color= "white"))
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2022 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php

- Scroll down and select the LWSP link next to Durham Municipality.
- Note the web address: https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
NCDEQs <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022")
print(NCDEQs)
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

3. The data we want to collect are listed below:

- From the "1. System Information" section:

- Water system name

- PWSID

- Ownership

- From the "3. Water Supply Sources" section:

- Maximum Day Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four separate variables.

> HINT: The first value should be "Durham", the second "03-32-010", the third "Municipality", and the last should be a vector of 12 numeric values (represented as strings), with the first value being "27.6400".

```
#3
water.system.name <- NCDEQs %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

PWSID <- NCDEQs %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

ownership <- NCDEQs %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

max.withdrawals.mgd <- NCDEQs %>%
  html_nodes("th~ td+ td , th~ td+ td") %>%
  html_text()
```

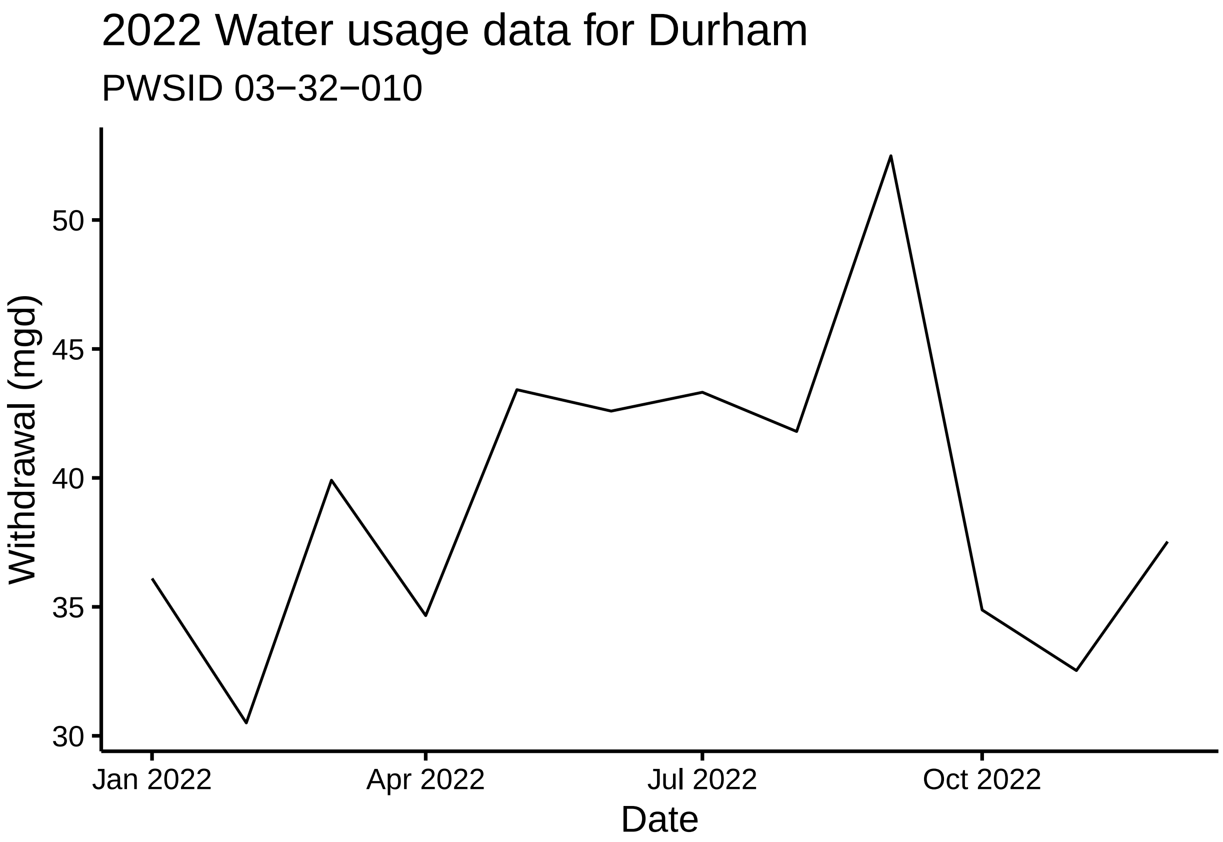
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   TIP: Use `rep()` to repeat a value when creating a dataframe.

   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the average daily withdrawals across the months for 2022

```
#4
Months <- NCDEQs %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()
the_year <- 2022
Date <- my(paste0(Months,"-",the_year))

df_withdrawals <- data.frame("Month" = Months,
                             "Date" = Date,
                             "Ownership" = ownership,
                             "PWSID" = PWSID,
                             "Year"= rep(2022, 12),
                             "Water System Name" = water.system.name,
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))


#5
ggplot(df_withdrawals,aes(x= Date, y= Max_Withdrawals_mgd)) +
  geom_line() +
  labs(title = paste("2022 Water usage data for",water.system.name),
      subtitle = paste("PWSID", PWSID),
       y="Withdrawal (mgd)",
       x="Date") +
  Aileentheme
```



6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

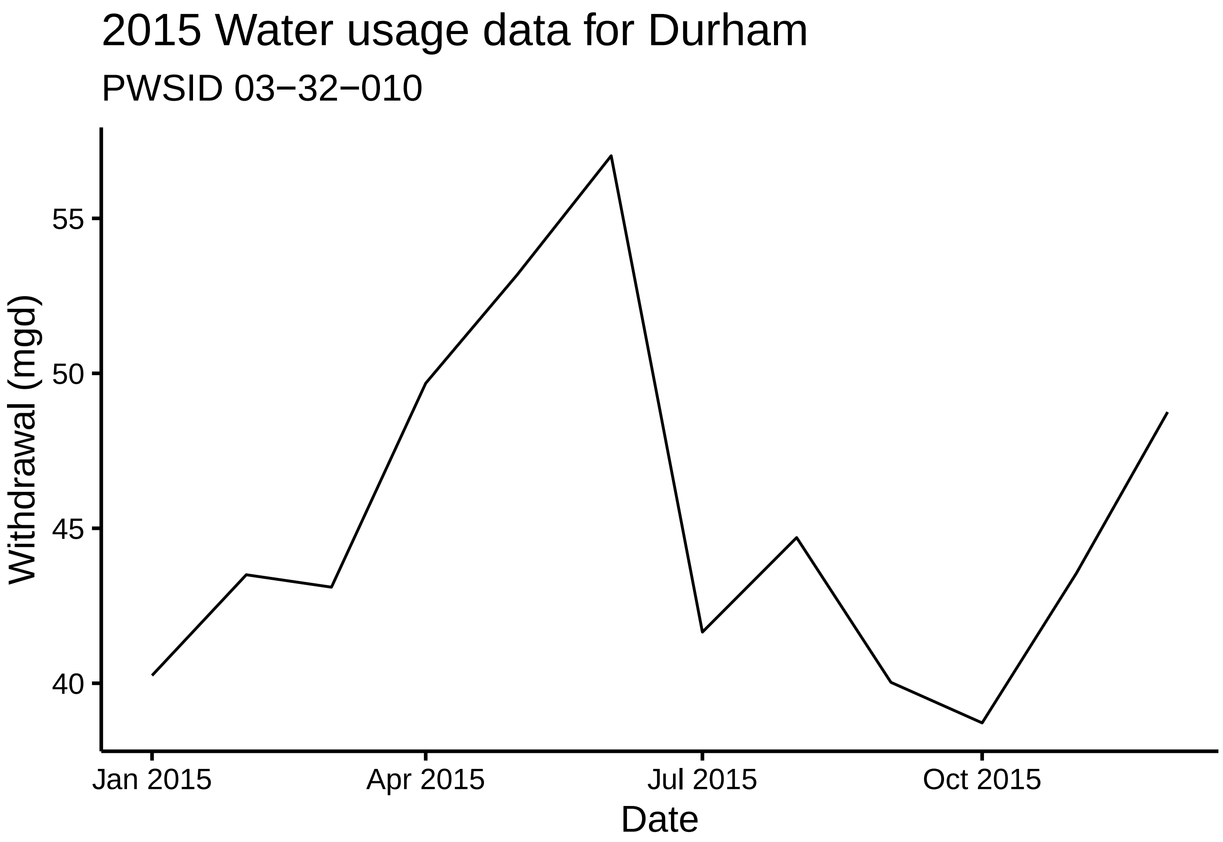
```
#6.
scrape.it <- function(PWSID, year){
#Retrieve the website contents
Website <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', PWSID, '&year=', year)
print(Website)
url <- read_html(Website)

#scrape the data frames
  water.system.name <- url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

PWSID <- url %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

ownership <- url %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

max.withdrawals.mgd <- url %>%
  html_nodes("th~ td+ td , th~ td+ td") %>%
  html_text()

#Convert to a dataframe
Month <- c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

df_YearPW <- data.frame("Month" = Months,
                        "Date" = my(paste0(Months,"-",year)),
                        "Ownership" = ownership,
                        "PWSID" = PWSID,
                        "Water System Name" = water.system.name,
                        "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

#Return the dataframe
return(df_YearPW)
}
```

7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
df2015 <- scrape.it("03-32-010","2015")
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2015"
```

```
ggplot(df2015,aes(x= Date, y= Max_Withdrawals_mgd)) +
  geom_line() +
  labs(title = paste("2015 Water usage data for",water.system.name),
       subtitle = paste("PWSID", PWSID),
       y="Withdrawal (mgd)",
       x="Date") +
  Aileentheme
```

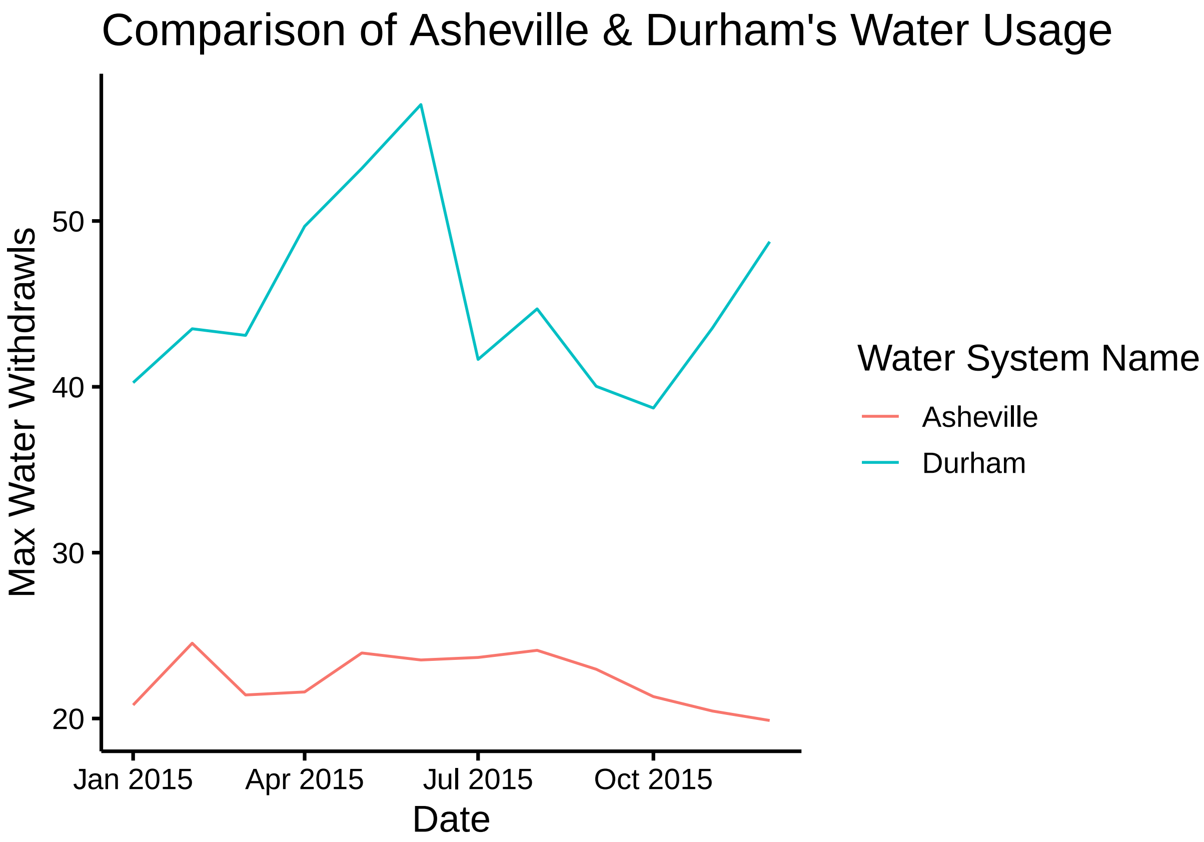


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
dfasheville2015 <- scrape.it("01-11-010","2015")
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
```

```
combined_df <- rbind(dfasheville2015,df2015)

plot_2 <- combined_df %>%
  ggplot(aes(x = Date, y =Max_Withdrawals_mgd, color=Water.System.Name)) +
  geom_line() +
  labs(
  x = "Date",
  y = "Max Water Withdrawls",
  title = "Comparison of Asheville & Durham's Water Usage",
  color= "Water System Name") +
  Aileentheme

plot_2
```



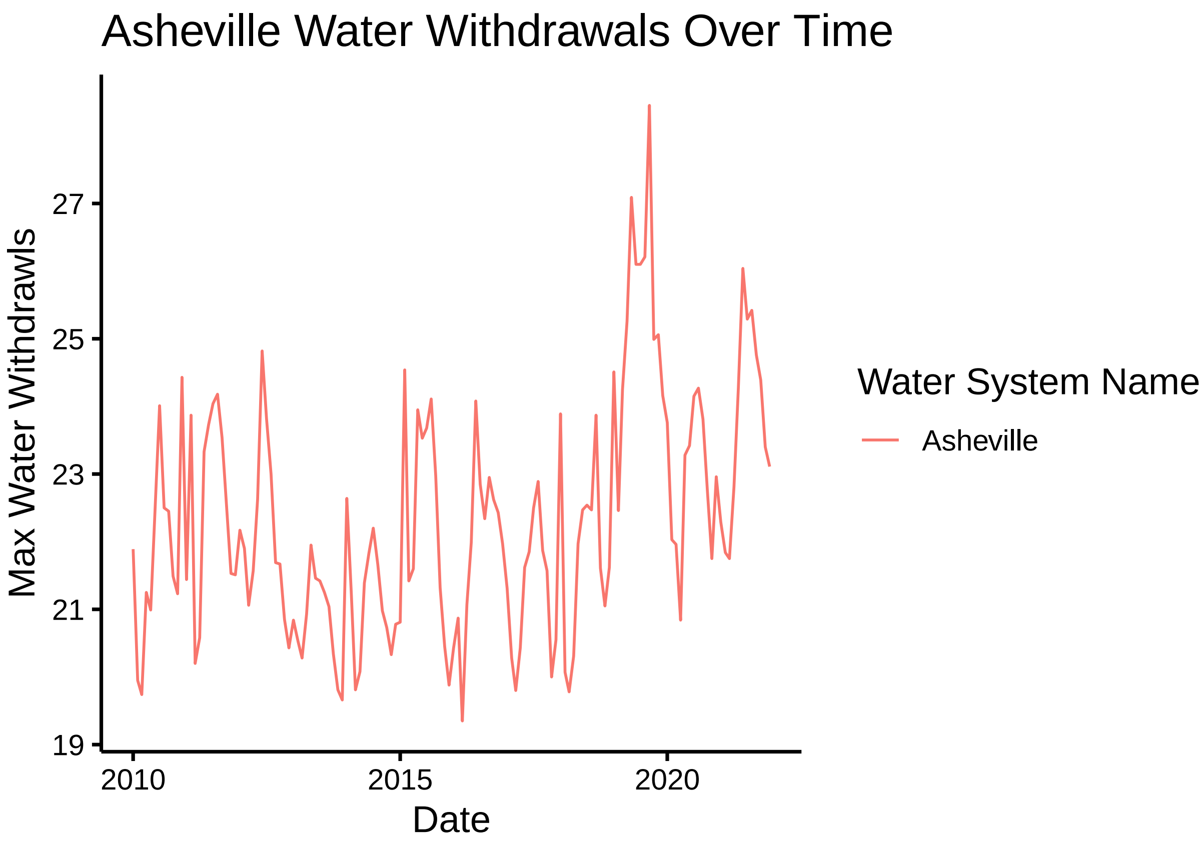
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
#Create a subset of years
All_Years = c(2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020,2021)
my_PWSID = "01-11-010"

#Use lapply to apply the scrape function
Asheville_dfs <- lapply(X = All_Years,
                  FUN = scrape.it,
                  PWSID = my_PWSID)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2010"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2011"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2012"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2013"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2014"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2016"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2017"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2020"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2021"
```

```
#Conflate the returned dataframes into a single dataframe
Asheville_dfs <- bind_rows(Asheville_dfs)

plot_3 <- Asheville_dfs %>%
  ggplot(aes(x = Date, y =Max_Withdrawals_mgd, color=Water.System.Name)) +
  geom_line() +
  labs(
  x = "Date",
  y = "Max Water Withdrawls",
  title = "Asheville Water Withdrawals Over Time",
  color= "Water System Name") +
  Aileentheme
plot_3
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? According to plot three, Asheville appears to have a higher water witdrawl in the later years (2019-2021) than the earlier years (2010-2015). There is a sike in water usage in 2019, followed by a sharp decrease in 2020, which migh be related to the Covid-19 pandemic.